DOCUMENT RESUME

ED 294 911 TM 011 621

AUTHOR Sympson, J. Bradford; Haladyna, Thomas M.
TITLE An Evaluation of "Polyweighting" in Domain-Referenced Testing.
SPONS AGENCY Navy Personnel Research and Development Center, San Diego, Calif.
PUB DATE Apr 88
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Software; Difficulty Level; Item Analysis; Latent Trait Theory; *Scoring; Test Items; *Test Reliability; Test Validity; *Weighted Scores
IDENTIFIERS *Domain Referenced Testing; Iterative Methods; *Polychotomous Scoring

ABSTRACT

A new approach to polychotomous scoring of test items, similar to "max-alpha" scaling (MAS) and known as polyweighting, has been developed. Unlike MAS, this new method of polychotomous scoring provides scoring weights for a given item that are independent of the difficulty of other items in the analysis. Moreover, the scoring weights are bounded so that an examinee can never receive more credit for an incorrect response than for a correct response. Polyweighting assigns a score weight to each item response category that is approximately equal to the mean percentile rank of examinees who chose that category in the item calibration sample. An iterative procedure is used to derive polyweights for a set of items. In a computer program written for this process, iterations continue until the mean squared correlation ratio between items and percentile ranks stops increasing. To evaluate the effect of polyweighting on the reliability and domain validity of test scores, data from 1,100 resident physicians completing a 200-item test on otolaryngology were obtained. Results indicate that polyweighting can provide a higher level of test reliability and domain-related validity than can traditional number/proportion-correct scoring. Further research should compare the effectiveness of polyweighting with item response theory methods. Seven tables and seven graphs are presented. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

BEST COPY AVAILABLE

ED294911

# An Evaluation of "Polyweighting" In Domain-Referenced Testing[1]

J. Bradford Sympson
Testing Systems Department
Navy Personnel Technical Center
San Diego, CA 92152

Thomas M. Haladyna
Education & Human Services Department
Arizona State University, West Campus
Phoenix, AZ 85017

Polychotomous scoring of test items, while not widely practiced, has a lengthy history. Haladyna and Sympson (1988) review that history and distinguish between two approaches to polychotomous scoring. One approach involves the assignment of differential scoring weights to item response categories. In this approach to polychotomous scoring, the test score is a linear function of the examinee's item response vector.

One method of linear polychotomous scoring has the unique property of maximizing coefficient-$\alpha$ (Cronbach, 1951) for the set of items calibrated (Guttman, 1941; Lord, 1958). This scoring procedure has been referred to by various names, including *reciprocal averages scaling* (Horst, 1935), *optimal scaling* (Bock, 1960), and *dual scaling* (Nishisato, 1980). Since these names are not suggestive of the method's primary distinguishing characteristic, we refer to it as "max-alpha" (MA) scaling.

MA scaling has two drawbacks as an approach to polychotomous item scoring. First, the scoring weights that are derived for an item depend on the difficulty level of the other items that are calibrated at the same time. If an item is calibrated along with a set of easy items, the obtained scoring weights will be different than if the item were calibrated along with a set of difficult items. Second, in order to maximize $\alpha$, the MA method often assigns weights to wrong answers that exceed the weight assigned to the correct response.

The second approach to polychotomous item scoring discussed by Haladyna and Sympson has a shorter history. This approach derives from item response theory (IRT). IRT models for polychotomous calibration of multiple-choice items have been introduced within the past two decades (Bock, 1972; Samejima, 1979; Sympson, 1981, 1983, 1986; Thissen & Steinberg, 1984). In this approach to polychotomous item scoring, the test score is a nonlinear function of the examinee's item response vector.

If the set of items calibrated with a polychotomous IRT model is unidimensional, and if the chosen model fits the items, the model parameters for any one item will be independent of the parameters obtained for other items. However, if the assumed model is not correct, IRT item parameters are dependent on both the examinee population that is sampled and the set of items that is calibrated. IRT calibration methods require fairly large samples ($N \geq 1000$ per item) in order to provide stable results.

In this paper, we introduce a new approach to polychotomous scoring of test items. This approach is similar to MA scaling in some regards, but provides scoring weights for a given item that are independent of the difficulty of other items in the analysis. Moreover, the scoring weights are bounded so that an examinee can never receive more credit for an incorrect response than for a correct response.

---

[1] Paper presented in C. E. Davis (Chair), *New Developments in Polychotomous Item Scoring and Modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, April 6, 1988. The research described here was supported by the Navy Personnel Technical Center IR/IED Program, the Joint Services CAT-ASVAB Program, and the Navy-ASEE Summer Faculty Research Program. The opinions expressed are those of the authors, are not official, and do not necessarily reflect the views of the Navy Department. Programmers for this research effort were Norine Araracap and Diane Scharer.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. BRADFORD SYMPSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011 621

*Polyweighting: An Approach to Polychotomous Item Scoring*[2]

"Polyweighting" is a scoring procedure that uses a different scoring weight for each item response category. An examinee's "polyscore" is equal to the mean of the scoring weights of the categories chosen by the examinee. Polyweighting does not require the assumptions of IRT, and can be applied with smaller samples than are commonly required with IRT models. Polyweighting does require that item calibration be carried out with a random sample of examinees from the population of interest.

Unlike some scoring methods, polyweighting gives the examinee more credit for correct answers to difficult questions and less credit for correct answers to easy questions. Conversely, polyweighting penalizes the examinee more heavily for wrong answers to easy questions than for wrong answers to difficult questions. This may be contrasted with number/proportion-correct scoring and with scoring under the the 1-parameter and 2-parameter logistic IRT models. The latter scoring methods assign scores to examinees in a manner that renders the scores independent of the difficulty of the questions answered correctly or incorrectly (Birnbaum, 1968, p. 458).

In polyweighting, the scoring weight ("polyweight") assigned to each item response category is approximately equal to the mean percentile rank of examinees who chose that category in the item calibration sample. An iterative procedure must be used to derive polyweights for a set of items. The procedure is as follows:

(1) Each examinee in the calibration sample is assigned a provisional score equal to the examinee's proportion correct among items the examinee was administered. We assume that different examinees may have been administered different items during the item calibration, but that an adequate number of examinees (e.g., 100 or more) were administered each "set" of items. We also assume that item sets were assigned to examinees randomly, so that each "item-set group" is randomly equivalent to other examinee groups.

(2) Since proportion-correct (PC) scores for examinees who are administered different item sets are not directly comparable (due to variation in difficulties and other characteristics of the items administered), each examinee's PC score is converted to a percentile rank relative to those examinees who were administered the same item set. This is equivalent to an equipercentile equating of PC scores from different item sets (Angoff, 1971, p. 563). For each examinee, his/her percentile rank is the proportion of examinees, among those who were administered the same item set, who obtained a PC score that was less than or equal to the PC score obtained by the given examinee, times 100.

(3) For each item, the mean percentile rank among examinees who chose each possible response category is determined. This computation includes all examinees who were administered a given item, even if they were administered different item-sets. At this point, if the mean percentile rank among examinees who chose the correct answer for a given item is less than the mean percentile rank among all examinees who were administered the item, the item is deleted from the analysis. This is equivalent to deleting an item if the *point-biserial* correlation between the correct answer and examinee percentile ranks becomes negative (Henrysson, 1971, p. 142).

(4) Next, for all items and all response categories, provisional polyweights are obtained by rounding the mean percentile rank among examinees who chose a category to the nearest integer. Since examinee percentile ranks range from a minimum possible value of $100(1/N)$ to a maximum possible value of 100, the provisional polyweights can range from 0 to 100. For a given item, if the provisional polyweight for an incorrect response is found to equal or exceed the provisional polyweight for

[2] The polyweighting procedure, described in this section, and the program POLY, described in the next section, were developed by the first author.

the correct response, the polyweight for the incorrect response is set equal to 1 less than the polyweight for the correct response. Thus, under polyweighting, an examinee can never receive more credit for an incorrect answer than for a correct answer.

(5) Given the provisional polyweights for all response categories, provisional examinee polyscores are computed. As stated earlier, an examinee's polyscore is equal to the mean of the polyweights of the categories chosen by the examinee. Since polyscores, like all raw test scores, are not comparable between examinees who have taken different item sets, the provisional polyscores are converted to percentile ranks within each group of examinees who have been administered the same set of items.

(6) Given the new percentile ranks for all examinees, the iterative procedure returns to Step 3, above. Steps 3 through 5 are repeated until a convergence criterion is satisfied. In the computer program POLY (Sympson, in preparation), iterations continue until the mean squared correlation ratio between items and percentile ranks stops increasing.

*An Example of Output from the Computer Program POLY*

Figures 1 through 4 show selected portions of the output from an item analysis using the computer program POLY. In this item analysis, polyweights were derived for 467 items that had been administered to 8141 applicants for military service. Each applicant was administered one of three 86-item vocabulary tests and one of six 35-item vocabulary tests. Thus, there were 18 different item sets administered to the examinees, with 121 items in each item set.

Figure 1 shows "Convergence Data" from the POLY run. Column 1 in Figure 1 gives iteration numbers. Iteration 0 is the iteration in which each examinee is assigned a raw score equal to his/her proportion correct among the items the examinee was administered. Subsequent iterations use provisional polyweights to compute examinee scores.

Column 2 in Figure 1 indicates how many items were included in the analysis during each iteration. In this example, one item was deleted, starting in iteration 2, because the point-biserial correlation between the item's correct response and percentile rank scores became negative.

Column 3 in Figure 1 gives the mean, over all retained items, of the squared correlation ratio between an item and percentile rank scores. In iteration 0, the value reported is the mean squared point-biserial correlation between correct responses and percentile ranks. For a given item, the squared point-biserial correlation for the correct response is equal to the proportion of variance in percentile ranks that is accounted for by knowing whether each examinee has selected the correct response. The point-biserial correlation is a widely-used index of item discriminating power when item scoring is dichotomous.

In subsequent iterations, the value reported in column 3 is the mean squared $\eta$ coefficient between an item and percentile ranks (Lord & Novick, 1968, p. 263). The squared $\eta$ coefficient between an item and percentile rank scores indicates the proportion of variance in percentile ranks that is accounted for by knowing which particular response category an examinee has selected. The $\eta$ coefficient is a more appropriate index of item discriminating power when items are scored polychotomously, and the weights assigned to categories are approximately equal to the mean test score of the examinees choosing each category. The $\eta$ coefficient for an item can never be smaller than the correct-answer point-biserial correlation. If there is any variation among the score means for the wrong-answer categories, the $\eta$ coefficient for an item will be larger than the point-biserial correlation.

Column 4 in Figure 1 shows the change ($\delta$) in the mean squared correlation ratio between iterations. This quantity serves as the convergence criterion in POLY runs. When $\delta$ becomes so small that it cannot be distinguished from zero, or if $\delta$ becomes negative, the iterations are terminated.

Column 5 in Figure 1 gives the mean value of coefficient-α in the analysis. Since there were 18 item sets in this example, each value reported in column 5 is the mean of 18 α coefficients. As the example shows, polyweighting increased the mean value of α for the item sets analyzed. The fact that scores based on polyweights have higher α coeffcients than do number/proportion-correct scores implies that test scores based on polyweighting will correlate more highly with domain scores and will have higher alternate-form reliabilities.

Column 6 in Figure 1 gives an index of "relative information." This index is based on the Spearman-Brown formula (Lord & Novick, 1968, p. 112). The Spearman-Brown formula gives the reliability of a lengthened test as a function of the initial reliability of the test and the proportionate increase in test length that is anticipated. However, rather than use the Spearman-Brown formula to predict reliability, one can rearrange the formula and use it to determine how much a given test would have to be increased in length in order to obtain a specified level of reliability (Nishisato, 1980, p. 118).

In POLY, the relative information index is set equal to 1.0000 in iteration 0. Subsequent to iteration 0, the formula used for computing relative information is:

$$h = \frac{\bar{\alpha}_p(1 - \bar{\alpha}_d)}{\bar{\alpha}_d(1 - \bar{\alpha}_p)} \quad , \tag{1}$$

where $\bar{\alpha}_d$ is the mean value of coefficient-α obtained under PC scoring (iteration 0) and $\bar{\alpha}_p$ is the mean value of coefficient-α obtained under polyweighting. This information index indicates the proportionate increase in test length that would be required in order to achieve the same reliability under PC scoring that has been achieved using polyweighting.

In the example shown in Figure 1, the POLY run terminated after iteration 3. At that time, the mean value of coefficient-α was .96140. When this value is substituted for $\bar{\alpha}_p$ in Equation 1, and the initial mean α of .94994 is substituted for $\bar{\alpha}_d$, the obtained final value of the relative information index is 1.3123. This indicates that a typical item set in this analysis would have to be increased in length by 31 percent (i.e., from 121 items to 159 items) in order to achieve the level of reliability under PC scoring that was achieved using polyweighting.

Figure 2 shows diagnostic information provided by POLY. In the example, Item 199 was deleted (not scored) starting in iteration 2, because the point-biserial correlation between the item's correct response and percentile rank scores was negative at the end of iteration 1. Items 53, 66, 99, 180, and 223 were scored, but have been flagged for special attention because each of these items had at least one incorrect answer with a positive point-biserial correlation that was larger than the point-biserial correlation for the correct answer.

Figure 3 shows two examples of the "Summary Item Analysis" provided by POLY for each item. Items 65 and 66 were selected as examples because Item 65 is quite easy and Item 66 is quite difficult. Moreover, Item 66 is one of the items flagged by POLY as needing special attention. Below the item-number for each item, a 25-character item-identification string is printed in parentheses. The user specifies this string for each item in the analysis. In Figure 3, both items came from Word Knowledge (WK) test-booklet number 1.

The columns headed "CAT." in Figure 3 contain response-category identification numbers. Category 0 is a pseudo-category that corresponds to "Not Administered." If an examinee's data-record contains a zero response-code for a given item, POLY does not use that item in computing the examinee's polyscore. In POLY, seven categories are available as scored catgories. The user must indicate the number of categories that are present for each item in the analysis. The number of categories can vary from item to item. In the examples in Figure 3, categories 1 through 5 correspond to choices "A" through "E" in these 5-alternative multiple-choice items.

In the examples in Figure 3, category 6 corresponds to "Omit." For a given item, a response-code of 6 was entered in the examinee data-record if the examinee did not answer the item, but he/she answered at least one item that appeared later in the same test booklet. Category 7 corresponds to "Not Reached." A response-code of 7 was entered in an

examinee's data-record if the examinee did not answer a given item, and he/she did not answer any subsequent items in the same test booklet. This use of categories 6 and 7 is specific to our example. During a POLY run, the last two categories for an item are not treated any differently than the other response categories (except category 0).

In Figure 3, the last value that appears in the columns headed "CAT." identifies a composite pseudo-category that collapses all actual response categories into one. In the examples, this pseudo-category is labeled "1-7" because items 65 and 66 were each specified to have seven categories. Summary statistics derived from all examinees who were administered a given item (i.e., from all categories other than category 0) are associated with this pseudo-category.

The entries in the columns headed "FREQ." in Figure 3 indicate how many examinees responded in each response category. The frequency shown for category 0 is the number of examinees who were not administered the item ($N$ = 5312 for items 65 and 66). The frequency shown for the composite pseudo-category (category 1-7 in the example) is the number of examinees who were administered the item ($N$ = 2829 for items 65 and 66). The other frequencies in this column correspond to the categories indicated in column 1. The double-asterisk (**) that appears between columns 1 and 2 identifies the keyed (correct) response for each item.

In Figure 3, the entries in the columns headed "PROP." indicate the proportion of the examinee sample that was associated with each response category. In these columns, the proportions are based on the entire examinee sample ($N$ = 8141 in the example). The entries in the columns headed "ADJ. PROP." (column 7) are adjusted proportions. These proportions are based on just the examinees who were actually administered an item. Thus, for Item 65, .3414 of the examinee sample gave the keyed response (category 3, or "C"). However, since only .3475 of the sample was administered the item, the adjusted proportion for category 3 is .9823, indicating that Item 65 was quite easy. This may be contrasted with the adjusted proportion for the correct answer to Item 66 (.1421), which indicates that Item 66 was quite difficult.

The columns headed "%-ILE MEAN" and "%-ILE S.D." in Figure 3 give the means and standard deviations of percentile rank scores among examinees associated with each response category. Means and standard deviations are printed out for the pseudo-categories (categories 0 and, in this example, 1-7) so that the user can check for obvious violations of the requirement that each item be administered to a random sample from the examinee population. For each item, the means and standard deviations for the two pseudo-categories should be similar. If they are not, it suggests that randomly equivalent item-set groups were not achieved.

In the example, the mean percentile rank among the 2779 examinees selecting the correct response on Item 65 was 50.90. The mean percentile rank among individuals choosing a wrong answer on this item ranged from a high of 12.15 among the 8 individuals who chose category 4 ("D") to a low of .64 among the 7 individuals who did not reach the item (category 7).

The mean percentile rank among the 402 examinees who selected the correct response on Item 66 was 61.26. The mean percentile rank among individuals choosing a wrong answer on Item 66 ranged from a high of 53.62 among the 1787 individuals who chose category 5 ("E"), to a low of 1.60 among the 8 individuals who did not reach the item (category 7). Most of the category means for wrong-answers are substantially higher for Item 66 than for Item 65.

Final (iteration 3) polyweights for items 65 and 66 are shown in the columns labeled "Scoring Weight" in Figure 3. These are just the category means from iteration 2, rounded to the nearest integer. For both example items, none of the wrong-answer scoring weights exceeded the polyweight for the item's keyed response, so none of the wrong-answer weights had to be bounded. If any weight had been bounded (set equal to 1 less than the polyweight for the keyed answer), a single asterisk (*) would have appeared to the right of the bounded weight.

Consideration of the scoring weights shown in Figure 3 gives an indication of the impact of polyweighting. An examinee who answers Item 66 correctly will receive more credit than an examinee who answers Item 65 correctly (61 vs. 51). Conversely, an examinee who answers Item 65 incorrectly will be penalized more heavily than an examinee who answers Item 66 incorrectly (a score of 12 or less if Item 65 is answered incorrectly vs. a score of at least 25 if Item 66 is reached but answered incorrectly).

The columns headed "R(C,%)" in Figure 3 contain unsquared point-biserial correlations between individual response categories and percentile rank scores. In the case of Item 65, there is only one positive point-biserial, the one for the keyed answer. In the case of Item 66, there are two positive point-biserials, one associated with the keyed answer, and one associated with category 5 ("E"). In fact, the point-biserial for category 5 is slightly higher than the point-biserial for the keyed answer. This is why Item 66 was mentioned in the output shown in Figure 2.

In Figure 3, the last summary statistic printed for each item is labeled "ETA(I,%)." This is the unsquared $\eta$ coefficient for the item. For Item 65, the $\eta$ coeffcient is only slightly larger than the point-biserial for the keyed response (.1975 vs. .1964), which indicates that polychotomous scoring of this item will add very little to measurement precision. This is not unexpected, since Item 65 is very easy and few wrong answers are observed. For Item 66, the $\eta$ coefficient is substantially larger than the point-biserial for the keyed response (.3449 vs. .1567), indicating that polychotomous scoring of this item will provide useful additional information about an individual's percentile rank within the examinee population.

An item such as Item 66 would often be discarded in a traditional item analysis because of the relatively large positive point-biserial correlation for category 5, a wrong answer. However, by using polyweighting, this apparently bad item can be retained and used to gather useful information about examinee ability. As indicated by the percentile mean for category 5, examinees who select this category are, on the average, of higher ability than examinees who select the other wrong answers. This fact is taken into account when the item is scored using the polyweights shown in Figure 3.

One would not want to use Item 66 in a test without further investigation of its psychometric properties and its content. To aid in this process, the program POLY can generate an "endorsement-rate table" for each item. The endorsement-rate table for Item 66 is shown in Figure 4. In order to create tables like the one shown in Figure 4, POLY can divide the examinee sample into as many as 100 ability groups, based on their percentile ranks. Then, using only the examinees who were administered a particular item, the proportion of examinees who gave each response is computed within each group. In the example, POLY was instructed to form 50 ability groups.

In Figure 4, the mean percentile rank within each ability group is shown in the column labled "%-ILE." The columns labled "CAT1" through "CAT7" contain the proportions selecting each category, within each ability group. It is notable that the proportion selecting the correct answer (category 4) is below chance-level (i.e., below .20) over virtually all of the ability range from the 1*st* to the 89*th* percentiles. Only in the top decile of this examinee sample does the proportion of examinees that select the keyed answer start increasing, with the highest ability group (the top 2%) selecting the keyed answer about 85% of the time.

In Figure 4, the endorsement rates for wrong answers usually decline as ability level increases, though the rate of decline varies between response categories. A pronounced exception to this pattern is observed for category 5, where the endorsement rate goes from chance-level, among the lowest ability groups, to over .80 within ability groups near the 75*th* percentile. In the top decile of the examinee sample, the endorsement rate for category 5 finally starts dropping, declining to about .15 in the highest ability group.

To aid interpretation of the endorsement-rate table shown in Figure 4, Figures 5 through 11 show graphic plots of the category endorsement rates for Item 66. The computer program POLY does not provide this type of plot, but such plots can be generated using the

endorsement-rate tables that are available from POLY. The plots in Figures 5 through 11 also show fitted functions that smooth and interpolate the plotted endorsement rates. It is clear from these plots that few examinees omit or do not reach Item 66 (categories 6 and 7), that most examinees select category 5, and that only the highest-ability examinees are not attracted to category 5.

The most important step in the evaluation of an unusual item is a careful inspection of its content. Some information about Item 66 may be of interest. First, as mentioned previously, Item 66 is a 5-alternative multiple-choice vocabulary item. This item asks the examinee to select the best synonym for "respite." The keyed answer is "rest" (category 4). Category 5, the very popular wrong answer, is "grudge." It appears that all but the most knowledgeable examinees were fooled by the presence of the word "spite" within the item stem. Further inspection of Item 66 gave no indication of a problem with the content of the item, so its use in a polychotomously-scored test seems appropriate.

More output from the example POLY run is shown in Figure 12. As mentioned earlier, there were 18 item sets administered in this item calibration example. Figure 12 shows summary statistics that were printed out for one of these item sets. Statistics provided by POLY include the number of examinees who were administered the item set, the number of items in the item set, the mean and standard deviation of raw polyscores for the item set, the minimum and maximum raw polyscore observed, the minimum and maximum standardized polyscore, and coefficient-$\alpha$ for the item set. In addition to these summary statistics, a table giving the mean raw polyscore and the mean standardized polyscore for each of 25 equal-frequency score groups is printed. This table allows the user to gain an impression of the shape of the distribution of raw/standardized polyscores for each item set. In Figure 12, it is clear that the distribution of raw/standardized polyscores was left-skew for item-set number 1.

*An Evaluation of Polyweighting*[3]

In order to evaluate the effect of polyweighting on the reliability and domain validity of test scores, data from 1100 resident physicians that had completed a 200-item test in the field of otolaryngology (the diagnosis and treatment of ear, nose, and throat disorders) were obtained. Five-hundred of these physicians were selected at random to make up "Sample A." Five-hundred different physicians were selected at random to make up "Sample B." The program POLY was then applied to the Sample A data in order to obtain summary item statistics and polyweights for all 200 items.

Next, using the set of 200 items as an item bank, 20 different assessment tests were assembled and applied in Sample B. These tests were assembled as follows:

(1) Three randomly-selected item sets of size 10 were designated as tests R10-1, R10-2, and R10-3. Three samples of items were used in order to obtain an indication of the amount of sampling variation in reliability and domain validity that could be expected when tests are assembled by randomly sampling items. Since items in the 200-item test had been allocated to 5 content categories by expert (physician) consultants, 2 items were randomly selected from each content category, in order to insure that each test was content valid.

(2) In a manner similar to the 10-item tests, 3 randomly-selected item sets of size 20 were designated as tests R20-1, R20-2, and R20-3. Each of these tests included items from one of the randomly-assembled 10-item tests. R20-1 included the items making up test R10-1, R20-2 included the items in test R10-2, and R20-3 included the items in test R10-3. In these tests, 4 items were randomly selected from each of the 5 content categories.

[3] The study described in this section was conceived and executed by both authors.

(3) Three randomly-selected item sets of size 30 were designated as tests R30-1, R30-2, and R30-3. Each of these tests included items from one of the randomly-assembled 20-item tests. R30-1 included the items making up test R20-1, R30-2 included the items in test R20-2, and R30-3 included the items in test R20-3. In these tests, 6 items were randomly selected from each of the 5 content categories.

(4) Three randomly-selected item sets of size 40 were designated as tests R40-1, R40-2, and R40-3. Each of these tests included the items from one of the randomly-assembled 30-item tests. R40-1 included the items making up test R30-1, R40-2 included the items in test R30-2, and R40-3 included the items in test R30-3. In these tests, 8 items were randomly selected from each of the 5 content categories.

(5) Using the results of the Sample A 200-item POLY run, tests of length 10, 20, 30, and 40 items were assembled using "traditional" item selection criteria. In this test construction procedure, items were selected that had the highest correct-answer point-biserial correlations, subject to a requirement that all item difficulties (proportions correct) had to be within .10 of the mean item difficulty in the 200-item domain. The resulting tests were designated as tests T10, T20, T30, and T40. Test T20 included the items making up test T10, test T30 included the items in test T20, and test T40 included the items in test T30. As before, item selection was accomplished within the designated content categories, with $k$ items being selected from each category for a $5k$-item test.

(6) Using the results of the Sample A 200-item POLY run, tests of length 10, 20, 30, and 40 items were assembled by selecting the items within each content category that had the highest $\eta$ coefficients. These four tests were designated as tests EM10, EM20, EM30, and EM40. Test EM20 included the items making up test EM10, test EM30 included the items in test EM20, and test EM40 included the items in test EM30. As before, $k$ items were selected from each content category for a $5k$-item test.

Each of the 20 tests described above was scored two different ways in Sample B. First, each test was scored by assigning a weight of 1 to all correct-response categories, a weight of 0 to all incorrect-response categories, and computing the mean weight among the categories selected. This gave the traditional PC score. Next, each test was scored using the polyweights derived in Sample A. For each Sample B examinee, his/her polyscore was the mean Sample A polyweight among the categories selected by the examinee.

For each of the 20 tests, Sample B item and test scores were used to compute coefficient-$\alpha$ for both PC scoring and for polyweighting. The two resulting values of $\alpha$ for each test were then used to compute a value of the relative information index defined in Equation 1.

Next, for each of the 20 tests, Sample B test scores and Sample B domain scores (based on all 200-items) were used to compute domain validities. For PC scoring, each examinee's domain score was the examinee's proportion correct on the 200-item test. For polyweighting, examinee domain scores were obtained by running POLY on the Sample B data for all 200 items. It is relevant to note that under PC scoring the weight (1 or 0) assigned to any given response category was the same whether the item appeared in a short assessment test or whether it was part of the domain. On the other hand, as a result of sampling error, the Sample A polyweight assigned to a response category during scoring of an assessment test in Sample B was, in general, somewhat different than the weight assigned to that category by POLY during the computation of Sample B domain scores.

Finally, after computing two Sample B domain validities for each test, the difference

$$d = \rho_p - \rho_d \quad , \tag{2}$$

where $\rho_p$ is the domain validity under polyweighting and $\rho_d$ is the domain validity under PC scoring, was obtained for each test.

*Results and Discussion*

Table 1 shows the results of this comparative evaluation of PC scoring and polyweighting. Inspection of Table 1 shows that for all 12 combinations of test length and test-construction method, polyweighting outperforms PC scoring in the cross-validation sample.

As expected, both coefficient-$\alpha$ and domain validity increase as test length increases, regardless of test-construction method and scoring method. Also, as might be expected, both coefficient-$\alpha$ and domain validity are higher for the systematically-constructed tests than for the randomly-assembled tests.

For all eight combinations of test-length and scoring method, tests made up of items with maximum $\eta$ coefficients are more reliable than tests assembled using the traditional method. However, the "EM" tests are not always superior to the "T" tests when domain validity is the criterion.

The $h$ statistics in column 4 indicate that, on the average, polyweighting increased coefficient-$\alpha$ by an amount that corresponds to a 30% increase in test length. The largest increases are observed for the randomly-assembled tests, where the mean value of $h$ is 1.41. Smaller increases are observed for the systematically-constructed tests, where the mean $h$ is 1.14. There is an indication that the "EM" tests benefit slightly more from polyweighting, since the mean $h$ for these four tests is 1.16, vs. 1.12 for the four "T" tests.

The $d$ statistics in column 7 indicate that, on the average, polyweighting increased domain validity by .072. Again, the largest increases are observed for the randomly-assembled tests, where the mean value of $d$ is .087. For the systematically-constructed tests, the mean value of $d$ is .049. It is relevant to note that the domain validity of polyscores from test T30 exceeds the domain validity of PC scores from test T40 (.862 vs. .847). Similarly, the domain validity of polyscores from test EM30 exceeds the domain validity of PC scores from test EM40 (.863 vs. .841).

In summary, the results of this study indicate that polyweighting can provide higher levels of test reliability and domain-related validity than traditional number/proportion-correct scoring. This has been demonstrated in the context of a cross-validation design. Further research is needed in order to compare the effectiveness of polyweighting with that of IRT scoring methods. Polyweighting should prove to be particularly useful when item calibration samples are small and/or the item domain of interest is multidimensional.

## References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Bock, R. D. (1960). *Methods and applications of optimal scaling* (Research Memorandum 25). Chapel Hill, NC: Psychometric Laboratory, University of North Carolina.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *Prediction of personal adjustment* (Bulletin 48, pp. 321-348). New York: Social Science Research Council.

Haladyna, T. M., & Sympson, J. B. (1988, April). Empirically-based polychotomous scoring of multiple-choice test items: Historical overview. Paper presented in C. E. Davis (Chair), *New Developments in Polychotomous Item Scoring and Modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 6*, 369-374.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291-296.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: University of Toronto Press.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville: Department of Psychology, University of Tennessee.

Sympson, J. B. (1981, October). *A nominal model for IRT item calibration*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Millington, TN.

Sympson, J. B. (1983, June). *A new item response theory model for calibrating multiple-choice items*. Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1986, August). Extracting information from wrong answers in computerized adaptive testing. Paper presented in B. F. Green (Chair), *New Developments in Computerized Adaptive Testing*. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.

Sympson, J. B. (in preparation). POLY: A computer program for polychotomous item analysis and test scoring.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49*, 501-519.

**Table 1**
**Cross-validated Reliability and Domain Validity of Proportion-correct Scores and Polyscores for Twenty Tests**

| Test | Reliability (α) | | | Domain Validity | | |
|---|---|---|---|---|---|---|
| | Type of Score | | | Type of Score | | |
| | PC | Poly | *h* | PC | Poly | *d* |
| R10-1 | .252 | .311 | 1.34 | .272 | .389 | .117 |
| R10-2 | .299 | .353 | 1.28 | .288 | .384 | .096 |
| R10-3 | .355 | .490 | 1.75 | .437 | .539 | .102 |
| R20-1 | .517 | .572 | 1.25 | .531 | .635 | .104 |
| R20-2 | .534 | .597 | 1.29 | .560 | .629 | .069 |
| R20-3 | .508 | .639 | 1.71 | .577 | .701 | .124 |
| R30-1 | .647 | .697 | 1.25 | .690 | .748 | .058 |
| R30-2 | .582 | .646 | 1.31 | .634 | .703 | .069 |
| R30-3 | .599 | .703 | 1.59 | .658 | .766 | .108 |
| R40-1 | .701 | .756 | 1.32 | .758 | .817 | .059 |
| R40-2 | .675 | .727 | 1.28 | .731 | .793 | .062 |
| R40-3 | .701 | .785 | 1.56 | .755 | .830 | .075 |
| T10 | .583 | .606 | 1.10 | .597 | .655 | .058 |
| T20 | .720 | .741 | 1.11 | .751 | .804 | .053 |
| T30 | .778 | .799 | 1.14 | .815 | .862 | .047 |
| T40 | .824 | .841 | 1.14 | .847 | .891 | .044 |
| EM10 | .625 | .657 | 1.15 | .606 | .653 | .047 |
| EM20 | .738 | .767 | 1.17 | .760 | .804 | .044 |
| EM30 | .810 | .832 | 1.16 | .815 | .863 | .048 |
| EM40 | .843 | .862 | 1.16 | .841 | .895 | .054 |

Figure 1

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141

CONVERGENCE DATA

| ITERATION | NO. ITEMS | MEAN SQUARED CORREL. | DELTA | MEAN ALPHA | RELATIVE INFO. |
|---|---|---|---|---|---|
| 0 | 468 | .157546 | ------- | .94994 | 1.0000 |
| 1 | 468 | .170002 | 0.125e-01 | .96128 | 1.3082 |
| 2 | 467 | .170258 | 0.256e-03 | .96138 | 1.3117 |
| 3 | 467 | .170257 | -0.106e-05 | .96140 | 1.3123 |

MEAN SQUARED ETA(I,%) HAS CONVERGED FOR 467 ITEMS.
(DELTA .LE. ZERO)



REPRODUCED AT GOVERNMENT EXPENSE

Figure 2

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141

THE FOLLOWING 1 ITEM(S) WERE NOT SCORED BECAUSE THE POINT-BISERIAL CORRELATION FOR THE KEYED RESPONSE BECAME NEGATIVE:

| ITEM | REMOVED STARTING IN ITERATION |
|---|---|
| 199 | 2 |

CHECK THE ANSWER KEY AND/OR THE ITEM(S).

---

FOR EACH OF THE FOLLOWING 5 ITEM(S), THE KEYED RESPONSE DOES NOT HAVE THE HIGHEST POINT-BISERIAL CORRELATION WITH PERCENTILE SCORE:

53 66 99 180 223

CHECK THE ANSWER KEY AND/OR THE ITEM(S).



REPRODUCED AT GOVERNMENT EXPENSE

Figure 3

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141

ITEM 65

(WK, BOOK 1, ITEM 65 )

SUMMARY ITEM ANALYSIS
(** INDICATES KEYED RESPONSE)

| CAT. | | FREQ. | PROP. | SCORING WEIGHT | %-ILE MEAN | %-ILE S.D. | ADJ. PROP. | R(C,%) |
|---|---|---|---|---|---|---|---|---|
| 0 | | 5312 | 0.6525 | ------ | 50.15 | 28.86 | 0. | ------ |
| 1 | | 9 | 0.0011 | 2.00 | 2.29 | 1.47 | 0.0032 | -0.0936 |
| 2 | | 16 | 0.0020 | 10.00 | 9.98 | 17.71 | 0.0057 | -0.1049 |
| 3 | ** | 2779 | 0.3414 | 51.00 | 50.90 | 28.48 | 0.9823 | 0.1964 |
| 4 | | 8 | 0.0010 | 12.00 | 12.15 | 25.06 | 0.0028 | -0.0701 |
| 5 | | 7 | 0.0009 | 12.00 | 11.52 | 16.67 | 0.0025 | -0.0666 |
| 6 | | 3 | 0.0004 | 10.00 | 9.99 | 11.90 | 0.0011 | -0.0453 |
| 7 | | 7 | 0.0009 | 1.00 | 0.64 | 0.55 | 0.0025 | -0.0854 |
| 1-7 | | 2829 | 0.3475 | ------ | 50.14 | 28.87 | 1.0000 | ------ |

ETA(I,%) = 0.1975

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141

ITEM 66

(WK, BOOK 1, ITEM 66 )

SUMMARY ITEM ANALYSIS
(** INDICATES KEYED RESPONSE)

| CAT. | | FREQ. | PROP. | SCORING WEIGHT | %-ILE MEAN | %-ILE S.D. | ADJ. PROP. | R(C,%) |
|---|---|---|---|---|---|---|---|---|
| 0 | | 5312 | 0.6525 | ------ | 50.15 | 28.86 | 0. | ------ |
| 1 | | 334 | 0.0410 | 39.00 | 39.26 | 25.97 | 0.1181 | -0.1378 |
| 2 | | 108 | 0.0133 | 25.00 | 24.83 | 21.09 | 0.0382 | -0.1746 |
| 3 | | 179 | 0.0220 | 29.00 | 29.14 | 25.15 | 0.0633 | -0.1890 |
| 4 | ** | 402 | 0.0494 | 61.00 | 61.26 | 35.76 | 0.1421 | 0.1567 |
| 5 | | 1787 | 0.2195 | 54.00 | 53.62 | 25.58 | 0.6317 | 0.1578 |
| 6 | | 11 | 0.0014 | 34.00 | 34.10 | 26.45 | 0.0039 | -0.0347 |
| 7 | | 8 | 0.0010 | 2.00 | 1.60 | 2.58 | 0.0028 | -0.0895 |
| 1-7 | | 2829 | 0.3475 | ------ | 50.14 | 28.87 | 1.0000 | ------ |

ETA(I,%) = 0.3449

REPRODUCED AT GOVERNMENT EXPENSE

Figure 4

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141
ITEM 66 (WK, BOOK 1, ITEM 66 ) CATS=7 KEY=4 (X,I2,X,I4,2X,F6.2,7(2X,F7.5))

| GROUP | N | %-ILE | CAT1 | CAT2 | CAT3 | CAT4 | CAT5 | CAT6 | CAT7 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 54 | 1.06 | 0.27778 | 0.11111 | 0.18519 | 0.16667 | 0.11111 | 0.01852 | 0.12963 |
| 2 | 56 | 3.01 | 0.21429 | 0.12500 | 0.19643 | 0.16071 | 0.26786 | 0.03571 | 0. |
| 3 | 57 | 5.00 | 0.14035 | 0.14035 | 0.12281 | 0.26316 | 0.33333 | 0. | 0. |
| 4 | 56 | 7.00 | 0.17857 | 0.16071 | 0.21429 | 0.17857 | 0.26786 | 0. | 0. |
| 5 | 57 | 9.00 | 0.12281 | 0.08772 | 0.17544 | 0.15789 | 0.43860 | 0. | 0.01754 |
| 6 | 56 | 11.01 | 0.25000 | 0.07143 | 0.14286 | 0.16071 | 0.37500 | 0. | 0. |
| 7 | 55 | 12.97 | 0.12727 | 0.12727 | 0.16364 | 0.09091 | 0.49091 | 0. | 0. |
| 8 | 58 | 14.96 | 0.12069 | 0.12069 | 0.08621 | 0.15517 | 0.50000 | 0.01724 | 0. |
| 9 | 57 | 17.01 | 0.12281 | 0.07018 | 0.08772 | 0.08772 | 0.63158 | 0. | 0. |
| 10 | 57 | 19.01 | 0.22807 | 0.03509 | 0.12281 | 0.17544 | 0.43860 | 0. | 0. |
| 11 | 55 | 21.01 | 0.12727 | 0.03636 | 0.05455 | 0.10909 | 0.67273 | 0. | 0. |
| 12 | 56 | 22.97 | 0.12500 | 0.05357 | 0.12500 | 0.10714 | 0.58929 | 0. | 0. |
| 13 | 56 | 24.95 | 0.25000 | 0. | 0.16071 | 0.03571 | 0.55357 | 0. | 0. |
| 14 | 58 | 26.96 | 0.08621 | 0.03448 | 0.08621 | 0.06897 | 0.70690 | 0.01724 | 0. |
| 15 | 58 | 29.01 | 0.12069 | 0.06897 | 0.10345 | 0.12069 | 0.58621 | 0. | 0. |
| 16 | 55 | 31.00 | 0.09091 | 0.03636 | 0.07273 | 0.03636 | 0.74545 | 0.01818 | 0. |
| 17 | 59 | 33.03 | 0.16949 | 0.06780 | 0.08475 | 0.08475 | 0.59322 | 0. | 0. |
| 18 | 56 | 35.05 | 0.21429 | 0.01786 | 0. | 0.05357 | 0.71429 | 0. | 0. |
| 19 | 54 | 37.01 | 0.11111 | 0.03704 | 0.01852 | 0.05556 | 0.77778 | 0. | 0. |
| 20 | 57 | 38.99 | 0.19298 | 0.05263 | 0.03509 | 0.12281 | 0.57895 | 0.01754 | 0. |
| 21 | 58 | 41.01 | 0.08621 | 0.05172 | 0.06897 | 0.06897 | 0.72414 | 0. | 0. |
| 22 | 56 | 43.01 | 0.12500 | 0.03571 | 0.01786 | 0.08929 | 0.71429 | 0.01786 | 0. |
| 23 | 57 | 45.04 | 0.08772 | 0.03509 | 0.10526 | 0.05263 | 0.71930 | 0. | 0. |
| 24 | 55 | 47.01 | 0.10909 | 0.05455 | 0.07273 | 0.09091 | 0.67273 | 0. | 0. |
| 25 | 57 | 48.99 | 0.14035 | 0. | 0.05263 | 0.07018 | 0.73684 | 0. | 0. |
| 26 | 55 | 50.96 | 0.05455 | 0.01818 | 0.03636 | 0.07273 | 0.81818 | 0. | 0. |
| 27 | 59 | 52.96 | 0.11864 | 0.05085 | 0.06780 | 0.10169 | 0.66102 | 0. | 0. |
| 28 | 56 | 55.00 | 0.14286 | 0. | 0.03571 | 0.08929 | 0.73214 | 0. | 0. |
| 29 | 57 | 56.98 | 0.10526 | 0.01754 | 0. | 0.07018 | 0.80702 | 0. | 0. |
| 30 | 54 | 58.94 | 0.20370 | 0.01852 | 0. | 0.03704 | 0.74074 | 0. | 0. |
| 31 | 56 | 60.93 | 0.12500 | 0. | 0.03571 | 0.07143 | 0.76786 | 0. | 0. |
| 32 | 60 | 62.97 | 0.13333 | 0.03333 | 0. | 0.01667 | 0.81667 | 0. | 0. |
| 33 | 56 | 65.00 | 0.10714 | 0.03571 | 0.05357 | 0.01786 | 0.78571 | 0. | 0. |
| 34 | 56 | 66.98 | 0.17857 | 0.01786 | 0.03571 | 0.07143 | 0.66071 | 0.03571 | 0. |
| 35 | 57 | 68.97 | 0.12281 | 0.03509 | 0.03509 | 0.08772 | 0.71930 | 0. | 0. |
| 36 | 58 | 71.02 | 0.01724 | 0. | 0.01724 | 0.12069 | 0.84483 | 0. | 0. |
| 37 | 56 | 73.04 | 0.08929 | 0. | 0.07143 | 0.12500 | 0.71429 | 0. | 0. |
| 38 | 56 | 75.02 | 0.08929 | 0. | 0.01786 | 0.05357 | 0.83929 | 0. | 0. |
| 39 | 58 | 77.02 | 0.08621 | 0.03448 | 0.03448 | 0.05172 | 0.77586 | 0.01724 | 0. |
| 40 | 56 | 79.04 | 0.14286 | 0. | 0.01786 | 0.05357 | 0.78571 | 0. | 0. |
| 41 | 56 | 80.99 | 0.01786 | 0. | 0. | 0.10714 | 0.87500 | 0. | 0. |
| 42 | 56 | 82.99 | 0.05357 | 0. | 0. | 0.08929 | 0.85714 | 0. | 0. |
| 43 | 59 | 85.03 | 0.06780 | 0.01695 | 0.03390 | 0.05085 | 0.83051 | 0. | 0. |
| 44 | 55 | 87.05 | 0.01818 | 0. | 0.01818 | 0.14545 | 0.81818 | 0. | 0. |
| 45 | 57 | 89.02 | 0.07018 | 0. | 0. | 0.15789 | 0.77193 | 0. | 0. |
| 46 | 56 | 91.02 | 0.10714 | 0. | 0. | 0.30357 | 0.58929 | 0. | 0. |
| 47 | 56 | 93.00 | 0.03571 | 0. | 0.03571 | 0.35714 | 0.57143 | 0. | 0. |
| 48 | 56 | 95.00 | 0. | 0. | 0.05357 | 0.44643 | 0.50000 | 0. | 0. |
| 49 | 56 | 96.95 | 0.01786 | 0. | 0.01786 | 0.73214 | 0.23214 | 0. | 0. |
| 50 | 62 | 99.03 | 0. | 0. | 0. | 0.85484 | 0.14516 | 0. | 0. |



REPRODUCED AT GOVERNMENT EXPENSE

REPRODUCED AT GOVERNMENT EXPENSE

# WK, BOOK 1, ITEM 66.1

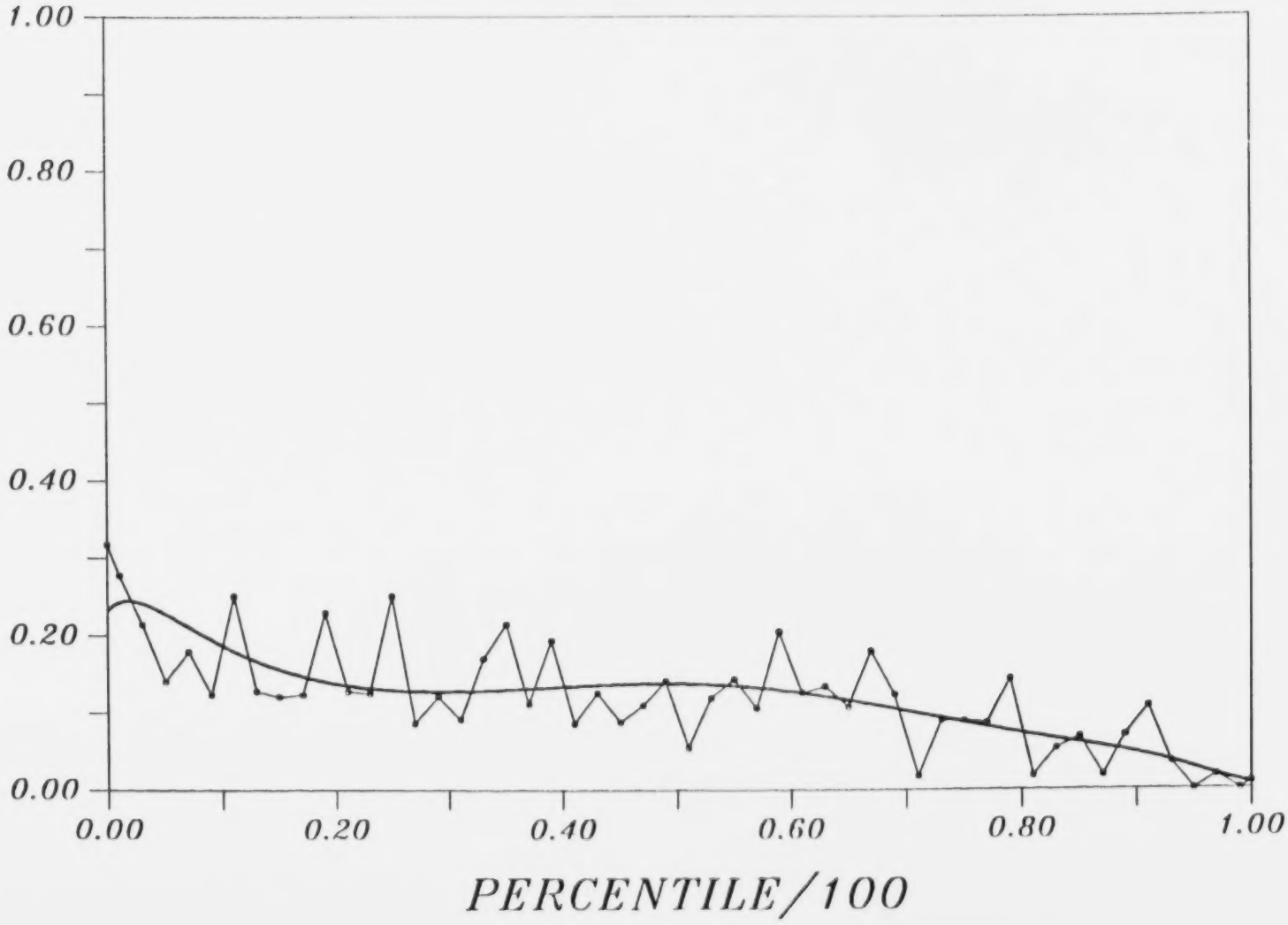


Figure 5

REPRODUCED AT GOVERNMENT EXPENSE

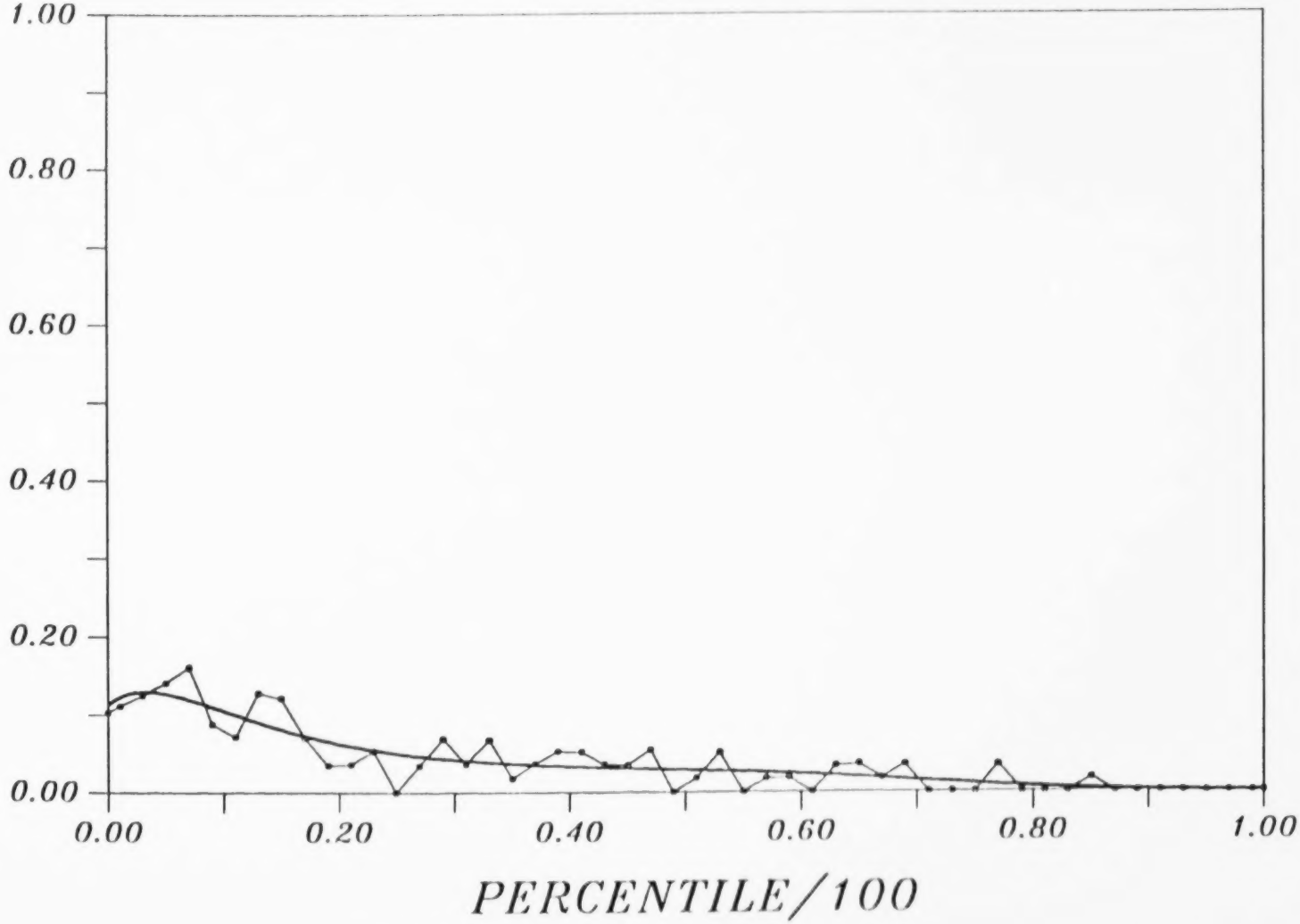


Figure 6

REPRODUCED AT GOVERNMENT EXPENSE

# WK, BOOK 1, ITEM 66.3

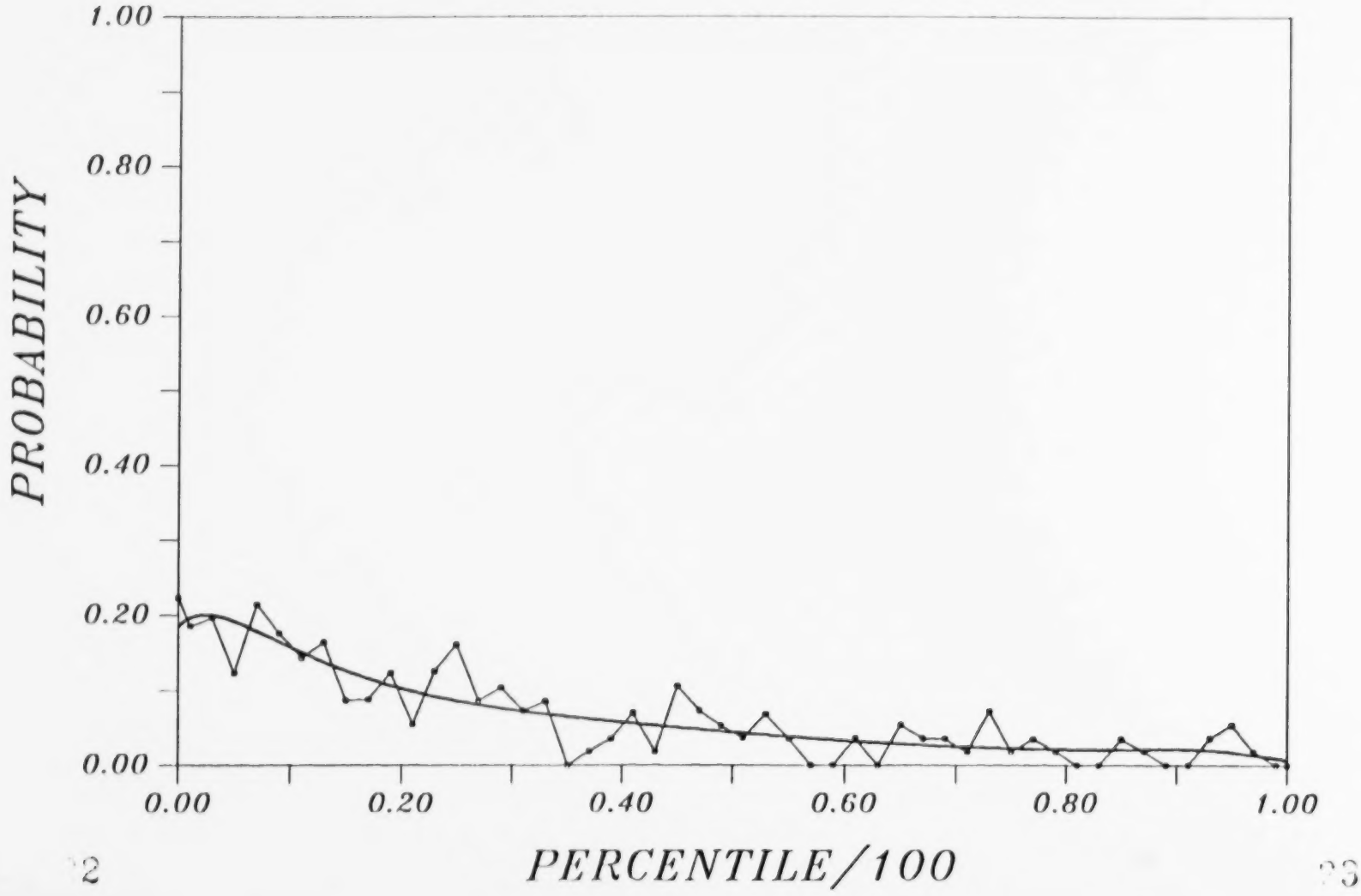


Figure 7

REPRODUCED AT GOVERNMENT EXPENSE



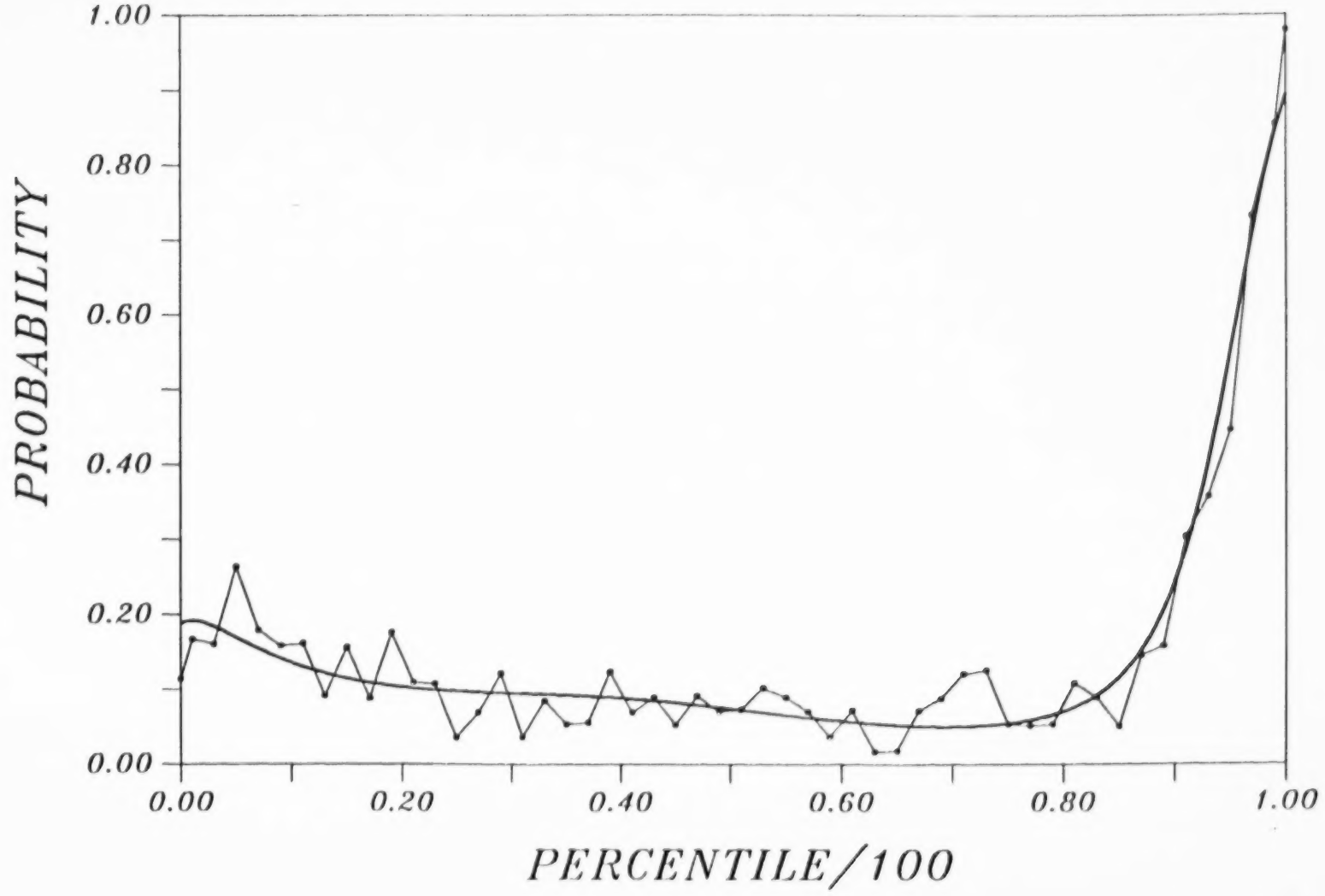


Figure 8

REPRODUCED AT GOVERNMENT EXPENSE

# WK, BOOK 1, ITEM 66.5

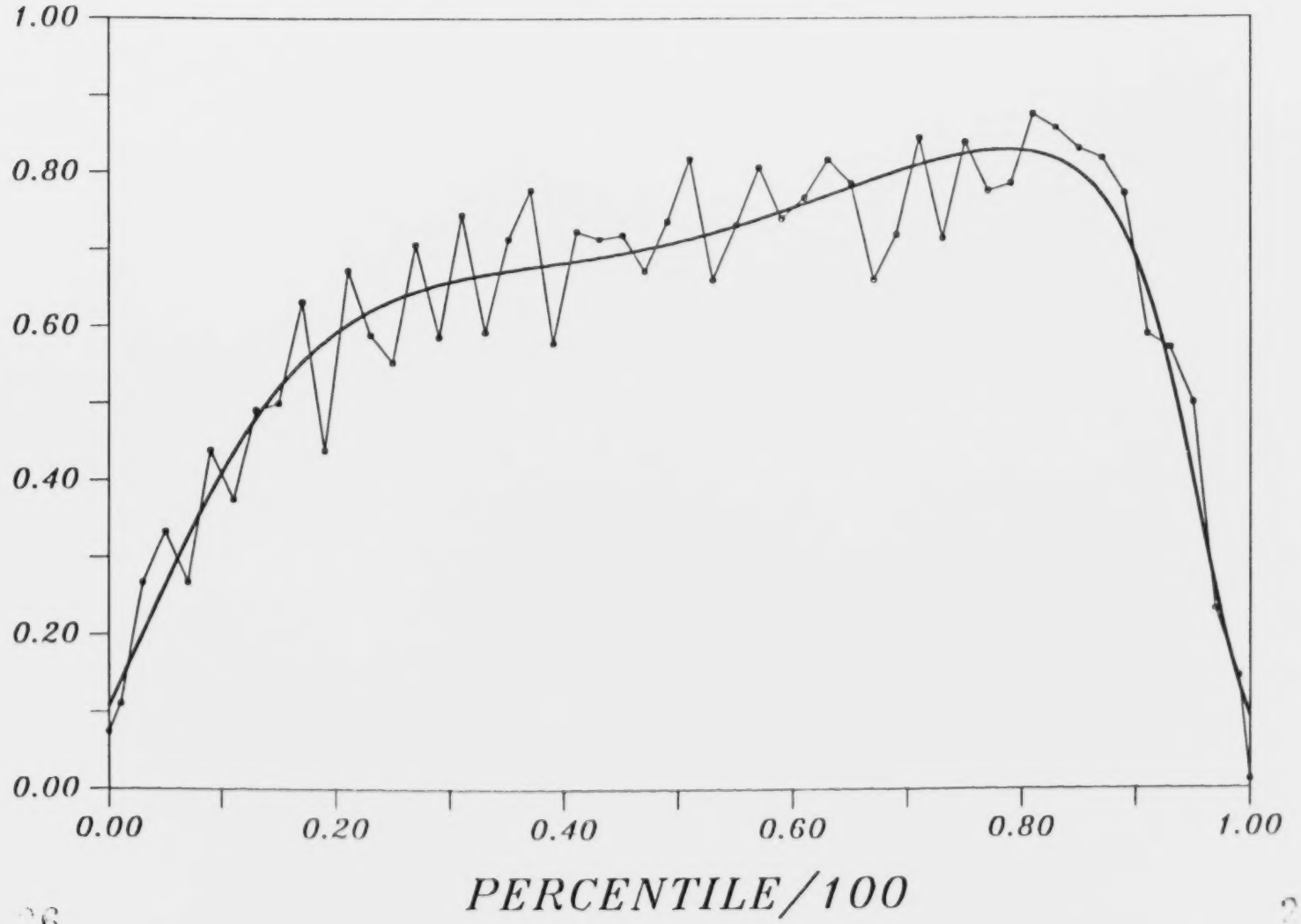


Figure 9

REPRODUCED AT GOVERNMENT EXPENSE



# WK, BOOK 1, ITEM 66.6

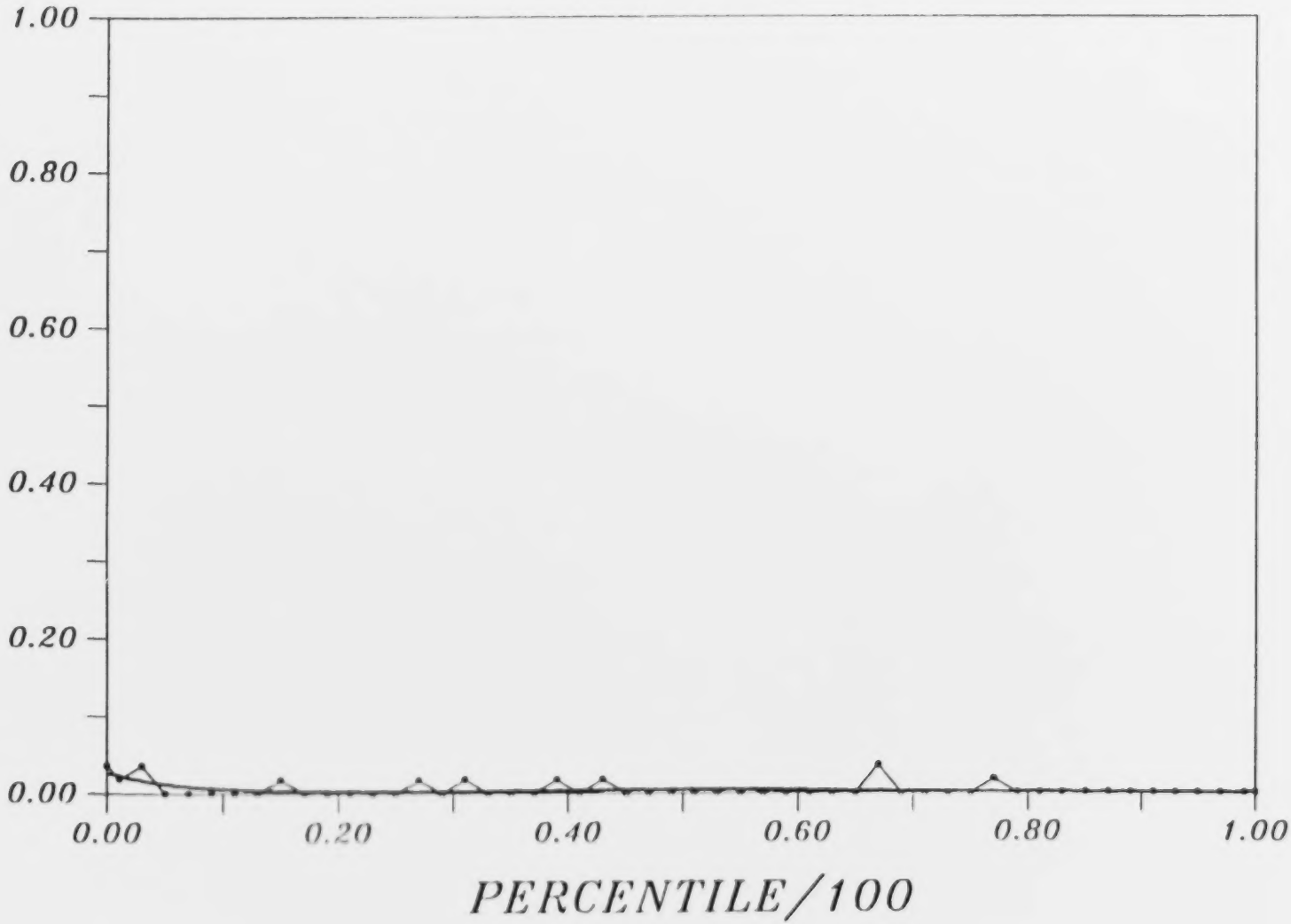


Figure 20

REPRODUCED AT GOVERNMENT EXPENSE

# WK, BOOK 1, ITEM 66.7

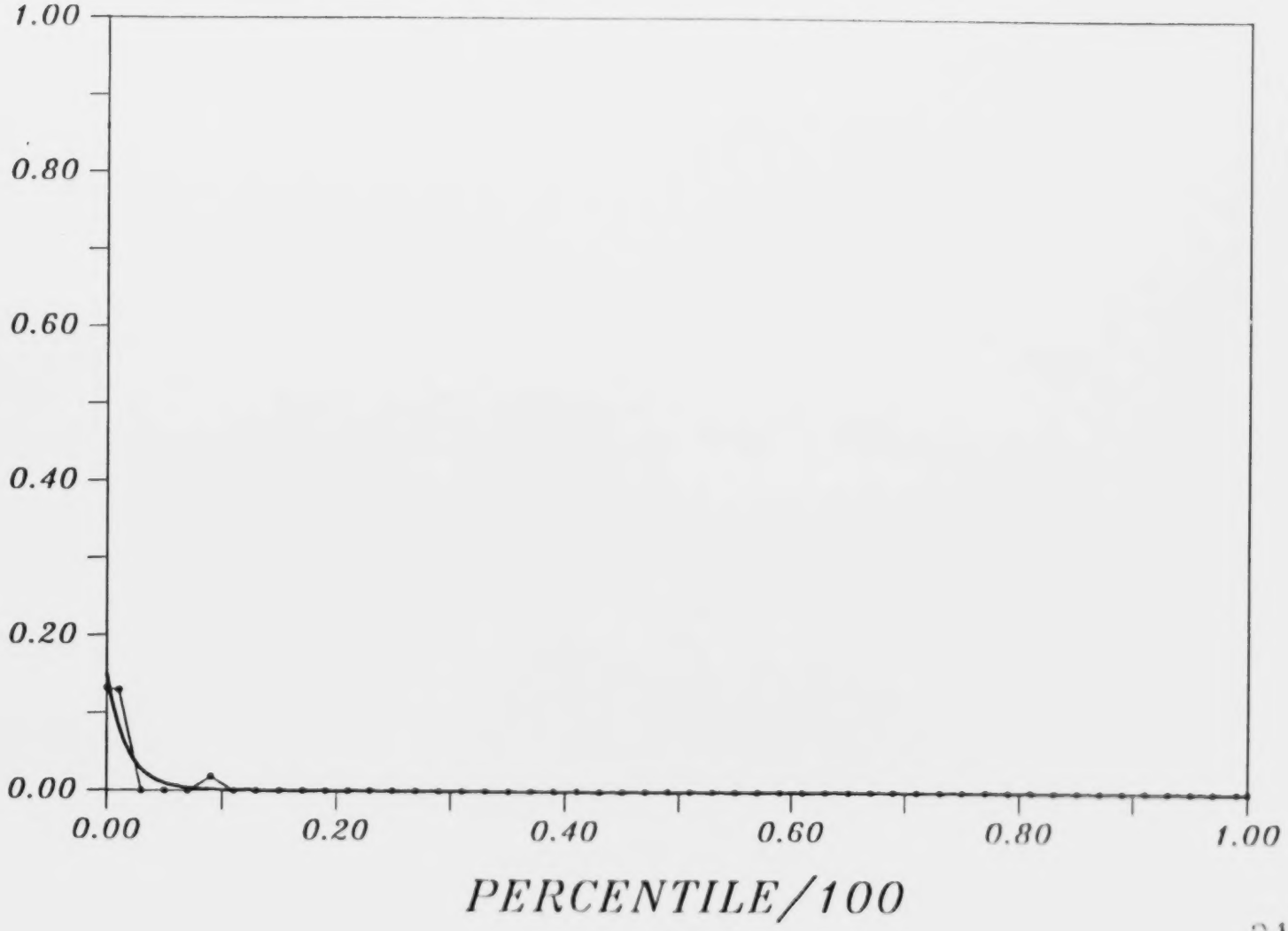


Figure 11

Figure 12

WORD KNOWLEDGE-ASVAB JOINT CALIBRATION, N=8141

SUMMARY STATISTICS FOR ITEM-SET 1

NUMBER OF EXAMINEES = 507
NUMBER OF ITEMS = 121

RAW SCORE MEAN = 49.95
RAW SCORE STANDARD DEVIATION = 4.44

MINIMUM RAW SCORE = 32.81
MINIMUM STANDARD SCORE = -3.8622
(CASE 5098)

MAXIMUM RAW SCORE = 56.38
MAXIMUM STANDARD SCORE = 1.4483
(CASE 358)

ALPHA = 0.95679

MEAN SCORES FOR ORDERED SCORE-GROUPS

| GROUP | N | RAW SC. | STD SC. |
|---|---|---|---|
| 1 | 21 | 37.40 | -2.8274 |
| 2 | 21 | 42.20 | -1.7476 |
| 3 | 21 | 44.13 | -1.3110 |
| 4 | 21 | 45.54 | -0.9945 |
| 5 | 21 | 46.52 | -0.7730 |
| 6 | 21 | 47.34 | -0.5894 |
| 7 | 21 | 47.98 | -0.4436 |
| 8 | 20 | 48.50 | -0.3277 |
| 9 | 20 | 48.82 | -0.2556 |
| 10 | 20 | 49.23 | -0.1631 |
| 11 | 20 | 49.86 | -0.0204 |
| 12 | 20 | 50.34 | 0.0883 |
| 13 | 20 | 50.84 | 0.2003 |
| 14 | 20 | 51.20 | 0.2822 |
| 15 | 20 | 51.55 | 0.3602 |
| 16 | 20 | 52.01 | 0.4633 |
| 17 | 20 | 52.37 | 0.5451 |
| 18 | 20 | 52.75 | 0.6295 |
| 19 | 20 | 53.13 | 0.7155 |
| 20 | 20 | 53.72 | 0.8494 |
| 21 | 20 | 54.16 | 0.9482 |
| 22 | 20 | 54.59 | 1.0448 |
| 23 | 20 | 54.99 | 1.1349 |
| 24 | 20 | 55.50 | 1.2511 |
| 25 | 20 | 56.05 | 1.3745 |

REPRODUCED AT GOVERNMENT EXPENSE